DOCUMENT RESUME

ED 065 591 TM 001 860

AUTHOR Edmonston, Leon P.; Randall, Robert S.
TITLE A Model for Estimating the Reliability and Validity of Criterion-Referenced Measures.
PUB DATE Apr 72
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Criterion Referenced Tests; Decision Making; *Evaluation Methods; Item Analysis; *Models; Statistical Analysis; *Test Reliability; *Test Validity

ABSTRACT

A decision model designed to determine the reliability and validity of criterion referenced measures (CRMs) is presented. General procedures which pertain to the model are discussed as to: Measures of relationship, Reliability, Validity (content, criterion-oriented, and construct validation), and Item Analysis. The decision model is presented in an appendix, the two sections of the model being Validation Procedures for Criterion Reference Measures: Unit Tests, and Validation Procedures for Criterion Referenced Measures: Mastery Tests. (DB)

ED 065591

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

A MODEL FOR ESTIMATING THE RELIABILITY AND VALIDITY OF CRITERION-REFERENCED MEASURES[1]

Leon P. Edmonston

and

Robert S. Randall

Southwest Educational Development Laboratory

Thomas Oakland

University of Texas

In a previous paper Oakland (1972) indicated that various theoretical models and statistical methods proposed for evaluating criterion-referenced measures (CRMs) were judged to be of limited use in validating CRM developed by the Southwest Educational Development Laboratory (SEDL). Problems inherent in these models as well as problems encountered in our own work (Edmonston, 1972) prompted an examination of alternatives which (1) would provide descriptive indices similar in philosophy to classical measures; (2) would base decision-making practices upon measures of student behaviors, teacher judgments, and data from objective instruments (including norm-referenced tests); (3) would provide comprehensible information to curriculum specialists untrained in evaluation procedures. The decision model presented in this paper is designed, in part, to meet these three objectives. The model currently is being used for validating CRM developed at SEDL.

A Decision Model

The objective information provided by CRMs is most important during the stage of formative evaluation. During this stage, data are utilized by curriculum developers and others to make judgments about how best to maximize the

TM 001 860

---

[1] Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April, 1972.

probability of learning an established set of objectives (Scriven, 1967); thus formative evaluation enables programs to undergo revisions and to introduce changes in a systematic process.

The primary function of an evaluation model during this period should be to provide decision-making procedures which utilize information for purposes of program revision. Information collected during this stage includes data from different types of CRMs, teacher observations, and to some extent, norm-referenced instruments. The decision model outlined in the Appendix provides a series of distinct steps for employing these data in order to make decisions about the psychometric properties of CRMs. The following discussion will be directed at the general procedures which pertain to this model; the various decision steps outlined in the model will not be examined separately.

Measures of Relationship

Criterion-referenced items usually are binary coded; the child either passes or fails to pass an item. Summaries of group performance at pre- and posttesting or at consecutive testings on a particular item can be displayed in a frequency table. Only the α X α table (α = 2) will be considered in this paper.

Consider the following table:

TABLE 1

A FOUR-CELL CLASSIFICATION TABLE OF PROPORTIONS OF STUDENTS WHO PASSED OR FAILED EACH OF TWO ITEMS

		Item 2: Pass	Item 2: Fail	Item 2: Total
Item 1	Pass	P11	P12	P1·
	Fail	P21	P22	P2·
	Total	P·1	P·2	1.0

Traditional statistics such as χ^2 and its derivatives provide tests of independence of the classification variables; while there is little question as to their adequacy as tests of independence, questions exist regarding their appropriateness as measures of association (e.g., Fisher, 1948, Goodman and Kruskal, 1954). Problems with employing χ^2 and $\emptyset$ with data from CRM have been discussed previously (Edmonston, 1972).

One alternative to these measures is to utilize the cell proportions themselves to provide information about the relationships between the classification variables. For example, a simple summation of the diagonal proportions, $\sum_{\alpha} \rho\alpha\alpha$, provides a very useful measure of agreement between categories. Procedures developed by SEDL for making decisions on the reliability and validity of CRMs essentially rely upon information displayed in tables like the one above. The coefficient of agreement, $\Sigma\ \rho\alpha\alpha$, is employed as one indicant of the relationship between cross classifications. As Goodman and Kruskal (1954) point out: "for the case in which the classes are ordered, but a meaningful metric is absent, we have been unable to find a measure [of reliability] better than $\Sigma\ \rho\alpha\alpha$" (p. 758).

A measure which provides information additional to $\Sigma\ \rho\alpha\alpha$ is a variance-free coefficient, Lambda (λ_r), based upon a probability model put forth by Goodman and Kruskal (1954). They define λ_r:

$$\lambda_r = \frac{\Sigma\ \rho\alpha\alpha - 1/2\ (PM\cdot + P\cdot M)}{1 \quad - 1/2\ (PM\cdot + P\cdot M)}$$

where $PM\cdot$ and $P\cdot M$ are the modal class frequencies for each of the two cross classifications. When no information exists as to how a method will classify an S randomly selected from the population[1], the probability of error in classification is $1-1/2\ (PM\cdot + P\cdot M)$. Going from the no information situation to a situation in which both classification methods are known, the probability of

---

[1] Goodman and Kruskal assume the population to be completely known in regard to the classifications. See Goodman and Kruskal (1963) for development of a sampling distribution as related to λ.

the error of classification decreases by $\Sigma \rho\alpha\alpha - 1/2(PM\cdot + P\cdot M)$. When normed by $1-1/2\ (PM\cdot + P\cdot M)$, λ_r may be interpreted as the relative reduction in the probability of error of classification when one goes from a no information situation to the other-method-known situation. The employment of λ_r in conjunction with $\Sigma\ \rho\alpha\alpha$ will be specified below.

Reliability

The decision model is formulated to serve as a flowchart for decision making. Step one in the model makes the determination as to whether a specified pass criterion has been attained initially on a particular item. If the criterion has been reached, reliability estimates are considered next. The question asked of the CRM performance data is the extent to which they fluctuate temporally; one means of acquiring this form of reliability estimate is the test-retest procedure. This is the only method of reliability investigated in the model. An example of this method and the probability models discussed above now will be described.

Eighty-three first-grade students were retested on a 28-item mastery test within a 10-day period. Data obtained from each item were arranged in the four-cell table described above (Table 1) and $\Sigma\ \rho\alpha\alpha$ was computed. Results are depicted in Table 2. Agreement ranged from .77 to 1.0 and most coefficients were above a minimally acceptable 85 percent standard.[1] λ_r also was computed. Results varied considerably, coefficients ranging from .00 to .75. These results also are presented in Table 2.

---

[1] This standard was arbitrarily established by the SEDL as the minimally acceptable standard for purposes of examining the reliability of its CRMs; others may choose to adjust the percent figure higher or lower.

TABLE 2

SOCIAL EDUCATION MASTERY TEST
RETEST RELIABILITIES
(N = 83)

Item	Frequencies PP	PF	FP	FF	$\rho\alpha\alpha$	λ_r*
1	71	0	8	4	.90	.50
2	58	16	2	7	.78	.44
3	71	9	2	1	.87	.15
4	76	3	0	4	.96	.73
5	59	11	4	9	.82	.55
6	57	12	7	7	.77	.42
7	80	3	0	0	.96	.00
8	77	3	1	2	.95	.50
9	69	5	2	7	.92	.67
10	64	5	9	5	.83	.42
11	74	4	1	4	.94	.62
12	79	1	2	1	.96	.40
13	73	2	3	5	.94	.67
14	66	1	6	10	.92	.74
15	67	8	1	7	.89	.61
16	63	4	5	11	.89	.71
17	66	6	3	8	.89	.64
18	79	1	2	1	.96	.40
19	73	5	5	0	.88	.00
20	62	8	5	8	.84	.55
21	65	8	3	7	.87	.56
22	73	1	3	6	.95	.75
23	71	2	6	4	.90	.50
24	74	2	2	5	.95	.71
25	83	0	0	0	1.00	Indeterminate
26	75	4	2	2	.93	.40
27	46	7	8	22	.82	.75
28	66	3	8	6	.87	.52

* The λ_r presented in this table = (λ_r + 1)/2. This has been done to have it range between .0 and + 1.0.

Minimally, CRM items should provide stable estimates of knowledge of curriculum content; $\Sigma\ p_{aa}$ and λ_r are employed to evaluate this stability. Concerning their use $\Sigma\ p_{aa}$ is used initially to judge the retest reliability (i.e., agreement) of each item. When an individual item has high retest agreement it is unnecessary to consider alternative measures for assessing reliability. However, when item reliability falls below an arbitrary criterion ($\Sigma\ p_{aa}$ = 85% is utilized by SEDL) and into a zone of decision (e.g., 76% - 84% agreement) in which additional information is needed to decide upon item retention or deletion, λ_r is introduced as a supplement to $\Sigma\ p_{aa}$.[1]

λ_r addresses questions somewhat different than the measure of agreement. It is employed as a descriptive measure of the amount of information (or error reduction) gained by employing a second item (the retest) in making curriculum or placement decisions. Because a primary function of CRM is assessment of group strength on a particular behavioral objective, if knowledge of the retest score provides additional information as to how students can be classified, the item is retained. Presently, no determination has been made at SEDL as to what constitutes an acceptable or minimal reduction in error for CRM.

---

[1] The model put forth by Goodman and Kruskal is considerably more embracing than the limited use to which it is being put here. It is especially important to recognize that λ_r is an index and not a number on any linear scale of equal units. For example, Goodman and Kruskal indicate that when complete independence exists between the cross classifications, λ_r assumes no particular value. This is not crucial because the measure is used only when dependence between classification variables is suspected.

It is also important to recognize that it is not always necessary to compute $\Sigma\ p_{aa}$. For example, if $\geq$ 90% pass the test and $\geq$ 90% also pass the retest, then it is unnecessary to undertake further computations because at least 80% will have passed both the test and the retest. Statistics do exist (e.g., Kappa) for determining whether the increase in the diagonal categories is greater than what would be expected when only the marginal frequencies are known.

Validity

The decision model provides for reliability estimation only when an item has reached a preestablished criterion level. If reliability is suitably high, then validity estimation is undertaken.

Content Validity. Content validation typically is seen as being central to validating achievement tests; this is also true for CRMs but with added emphasis. CRM items are sampled theoretically from a large item domain, and as such, must be representative of a specified behavioral objective or activity. Because typically one item only is employed to assess an objective, the need for adequate content validation procedures is strong. The decision model emphasizes content validation (i.e., IB and IVB) when objective data indicate possible item deficiencies. However, by definition, content validity techniques are employed first, during the curriculum writing and item construction period. This period is not covered directly within the decision model.

Criterion-Oriented Validity. A second method of instrument validation is the criterion-oriented approach. This includes both concurrent and predictive validity. Objective testing instruments are only one means of validity assessment; other measures might include behavior ratings by teachers and trained observers. In order to obtain complete information about an item and the objective which it assesses, the relationship of a CRM item to other measures are considered. The decision model provides for examining the relationship between CRM item performance and ratings by both teachers and trained observers as well as performance on suitable NRM items. These are all forms of concurrent validation (see IIIA, B and E; VIA, B and C in the model).

As indicated in another paper (Kennedy, 1972) tests of curriculum mastery represent the higher order concept taught within several curriculum units.

Consequently, performance on the unit test item representative of this concept should be predictive of mastery test performance. In addition, recognizing that a simplicial pattern (Guttman, 1955) often is found when trials on a learning task are intercorrelated, then unit test items which are more temporally proximate to mastery test items should agree more strongly with the mastery test items than unit test items sequenced earlier in the curriculum (see IIID of the model). This finding was early put into an ability framework by Woodrow (1938) who stressed that "final scores are dependent upon a different pattern of abilities than initial scores" (p. 277). This should be true especially when a 4-6 month period exists between unit and mastery tests.[1]

As a test for a simplicial pattern and of unit test validities, scores on unit items representing language identification of body parts were obtained from 40 three-year-old Ss receiving SEDL's Bilingual Early Childhood Program. Items from unit tests 2, 3, 7 and 10 were matched to their corresponding mastery test item; these units are sequenced numerically, approximately two weeks apart. Thus, unit 10 would be given approximately four months after unit 2. The coefficient of agreement was computed between each unit and mastery test item using a 2 X 2 table to estimate unit and mastery test validity.[2] Results are given in Table 3.

As predicted, a moderate trend exsists between these unit test items and the mastery test item; items from unit tests 3 and 7 exhibit little difference. The same procedure was undertaken for items measuring language identification of body senses. Results are presented in Table 4. As evidenced, the

---

[1] Naturally, Woodrow was referring to patterns obtained in a matrix of intercorrelations. The intent above is only to examine the relationship between unit and mastery test scores while referencing a familiar concept in the multivariate learning literature.

[2] When validity estimates are made, summation of off-diagonal cells can also be undertaken for equal predictive efficiency. This approach is not appropriate to answering the above questions.

TABLE 3

AGREEMENT INDICES BETWEEN UNIT TEST ITEMS
AND THEIR CORRESPONDING MASTERY ITEM

LANGUAGE IDENTIFICATION OF BODY PARTS

Unit Test	Item Number	Σ ραα
2	7	.73
3*	5, 6	.82
7	2	.81
10	10	.88

* Based on the summation of agreement indices for two highly similar items.

TABLE 4

AGREEMENT INDICES BETWEEN UNIT TEST ITEMS
AND THEIR CORRESPONDING MASTERY ITEM

LANGUAGE IDENTIFICATION OF BODY SENSES

Unit Test	Item Number	Σ ραα
1	3	.70
2	10	.76
3*	5, 6	.74
4	8	.43
7*	1, 2	.85
10	8	.68

* Based on the summation of agreement indices for two highly similar items.

expected simplicial pattern was not obtained. This would appear to indicate that the abilities necessary for attainment of objectives related to the concept early in the program are still important during the later units of the curriculum. It also indicates the low predictive validation of items taken from units four and ten. Of course, this hypothesis is quite tentative due to the limited number of data points considered as well as the absence of a more thorough design such as the incorporation of factorially pure reference tests. For purposes of data presentation, results for both reliability and validity can be displayed easily in a square symmetric matrix with reliabilities in the diagonals and validities as the off diagonal elements. See Table 5 for an example of this.

TABLE 5

SQUARE SYMMETRIC MATRIX WITH RELIABILITIES
IN THE DIAGONALS AND VALIDITIES
AS OFF-DIAGONAL ELEMENTS

$$A = \begin{bmatrix} a11 & a12 & \cdots & a1n \\ a21 & a22 & \cdots & a2n \\ \cdots & \cdots & \cdots & \cdots \\ am1 & am2 & \cdots & amn \end{bmatrix}$$

Construct Validation. Construct validity typically is employed in correlational research as an index of the extent to which a measure operationalizes or objectifies a theoretical construct. Opinions as to the meaning and nature of construct validation are too diverse to be considered here (e.g., Cronbach and Meehl, 1955; Bechtoldt, 1959). A methodological approach particularly appropriate to the construct validation of CRMs is advocated by Nunnally. He writes (1967) that the "measurement" and "validation" of constructs involves nothing more than the determination of an internal structure among a set of measures (e.g., all those items relating to the concept "size"), the development of cross structures between several sets (e.g., between items relating to "size" and those relating to "number"), and the consequent formation of "a network of probability statements" among measures within and between sets (p. 95).

Nunnally's approach is appropriate to correlational methods but his procedures are adapted easily to the measure of agreement when one is interested in determining whether mastery of one set of objectives is related to mastery of another set of objectives. One approach towards determining these internal structures is to relate CRMs to test items which already have been validated. For example, items on the Test of Basic Experiences: Social Studies were matched (using face validity) with unit test items in SEDL's Multicultural Social Education Program which appear to measure similar abilities. Σ pαα was determined for each set of items. Results are illustrated in Table 6.

TABLE 6

AGREEMENT INDICES (Σ ραα) BETWEEN MATCHED UNIT TEST AND TOBE ITEMS

TOBE ITEM	Σ ραα
First Grade	
"Mark the one who won't get hurt."	.55
"Mark the sad boy."	.83
"Mark what you do most in school."	.84
"Mark what your body needs."	.79
"Mark what you should wear for rain."	.75
"Mark the girl who is ready for a party."	.75
"Mark the boys who are doing the right thing."	.88
"Mark what the carpenter uses."	.64
Second Grade	
"Mark the boys who are doing the right thing."	.95
"Mark the sad boy."	.85
"Mark what your body needs."	.85

Six of the 11 items reached at least 83% on the agreement index while 9 of 11 reached 75%. Indices falling below the 80% level may indicate that the items were inadequately matched, or that one or both of the items failed to operationalize the concept under consideration adequately. If reliability and predictive validity procedures outlined in the decision model are followed, some initial judgement may be made as to why the relationship is not suitably high.

Early work in the area of determining a network of statements between sets of items operationalizing learning objectives was undertaken by Gagné and Paradise (1961). Their model relates the rate of completion and actual achievement of learning tasks to individual differences in both task relevant learning sets and abilities. Knowledge is viewed as a subset of "subordinate capabilities" or "learning sets" with learning sets being arranged hierarchically, ranging from basic broad abilities to more specific, subordinate learning sets in the hierarchy. In order to acquire knowledge essential for completing a task, a high degree of positive transfer from lower to higher order learning sets must result (theoretically, a Guttman-like prediction of 100% positive transfer).

Single items are employed to measure each learning set; the transfer measure is based upon the four possible pass/fail relationships like those displayed in Table 1. Gagne and Paradise thus bring in the sequential aspect to curriculum formation. The position outlined by Gagne and Paradise is similar to the approach touched upon by Roudabush and Green (1971). The specification of a hierarchy of learning sets among items would seem to be the ultimate goal of construct validation procedures enabling the development of internal and cross structures between items and the consequent understanding of the interrelatedness of all curriculum areas.

Item Analysis

Displaying CRM data in cross classification tables is appropriate not only for utilizing variance free measures but the information provided by individual cells can also be adapted directly to the philosophy underlying CRMs. For example, when a program of treatment intervenes between testings, the maximally discriminating CRM item is one which distinguishes students who fail at pretesting and pass at posttesting, and thus discriminates those who profit by the educational intervention. Carrying this to the extreme, if all students pass the posttest, typical item discrimination indices (e.g., comparing item performance of those in the upper and lower 27% of the total score distribution) would fail to discriminate adequately.

Cox and Vargas (1966) consider problems associated with item analysis procedures employed with CRMs; they compare a classical discrimination index (D) as described above and a posttest-pretest Difference Index (D_{pp}). Problems with their procedure have already been described (Oakland, 1972). Employing the notation in Table 1, their procedure undertakes the following:

$$(P11 + P21) - (P11 + P12) = P21 - P12$$

As evidenced, the Cox and Vargas procedure subtracts Ss who pass the pretest and fail the posttest from those who fail the pretest and pass the posttest. A less narrow refinement of their approach is to consider the off-diagonal cells separately. For example, consideration of P21 would indicate only those who passed a posttest and failed the pretest. Conversely, lease discriminating items would have high values in category P12. Appropriate criterion levels are easily established for each of these measures.

A P P E N D I X

Decision Model

VALIDATION PROCEDURES FOR CRITERION REFERENCE MEASURES: UNIT TESTS

I. Percent correct: it is desirable that ≥ 80%[1] of students who have received the curriculum should pass each unit test item.[2]

A. If ≥ 80% pass the item, go to II.

B. If < 80% pass the item, examine the item so as to discover possible deficiencies in the item (e.g., the item may be unrelated to the curriculum, the wording within the item may be misleading, or there may be errors in media).

(1) If these or other defects are apparent, momentarily discontinue further analyses of this item. Correct the apparent defects and readminister the items when it is possible to do so.

(2) If no defects are apparent, go to II.[3]

II. Determine the test-retest reliability for each item ((A+D)/N).

A. If test-retest reliability is suitably high, go to III.

B. If test-retest reliability is unsuitably low, examine the item and the conditions under which the item was administered. Continue further analyses on this item only after the item is corrected and readministered.

---

1 As there are no firm guidelines for selecting 80% as the acceptable criterion, this figure can be adjusted. A downward or upward adjustment in the criterion level does not influence appreciably other aspects of the proposed decision model.

2 The percent correct figure for children who have not received the curriculum should be roughly equal to 1/number of options in the item (i.e., chance).

3 Even though < 80% pass an item, the item may have suitable reliability, thus ruling out one factor which may account for the low percent correct figure.

	P	F
P	A_{40}	B_{10}
F	C_{10}	D_{40}

III. Validity: Unit Test Items

There should be a high level of agreement between the level of achievement children demonstrate in the classroom (and observed by classroom personnel) and their performance on test items measuring achievement.

A. There should be a high level of agreement[4] ((A+D)/N) between the achievement level of students as estimated by their teacher and their performance on a unit test item--provided that both refer to the same behavioral objective. The level of agreement will be determined for each unit test item.

(1) If (A+D)/N is suitably high, go to IIIB.

(2) If (A+D)/N is unsuitably low, try to determine possible reasons for the discrepancies in ratings and continue to IIIB.

B. There should be a high agreement ((A+D)/N) between the achievement level of students as estimated by trained observers and their performance on a unit test item--provided that both refer to the same behavioral objective. The level of agreement will be determined for each unit test item.

(1) If (A+D)/N is suitably high, go to IIIC and IIID.

(2) If (A+D)/N is unsuitably low, try to determine possible reasons for the discrepancies in ratings and continue to IIIC and IIID.

C. There should be a high level of agreement ((A+D)/N) between performance on each unit test item and performance on a mastery test item provided that both the unit test item and the mastery test item (1) are suitably reliable and (2) are designed to measure a common

---

4

Teacher Ratings	Item Performance: P	Item Performance: F
P	A	B
F	C	D

behavioral objective. The level of agreement between these items will be determined when appropriate.

D. Also, the level of agreement between a series of unit test items and a mastery test item may become increasingly higher as the unit test item more closely approximates the mastery test item. For example, consider the relationships ((A+D)/N) between seven individual unit test items and one mastery test item.

Unit Test Items	(A+D)/N	
1	.40	In general, this demonstrates a desirable relationship; however, items 3 and 6 appear to evidence deficiencies. If the reliability of item 3 and 6 is suitable, we would examine their validity estimates obtained from previous measures of validity.
2	.50	
3	.20	
4	.70	
5	.80	
6	.50	
7	.90	

However, consider another example.

Unit Test Items	(A+D)/N	
1	.40	Given these relationships, we may question the validity of the mastery test item provided that the unit test items demonstrate suitable reliability as well as suitable validity as determined from previous measures of validity (e.g., II, IIIA, and IIIB).
2	.30	
3	.20	
4	.30	
5	.60	
6	.40	
7	.10	

The level of agreement between these items will be determined when appropriate.

E. There should be a high level of agreement between performance on an item from a unit test and performance on an item from a NRM provided that the items from the unit test and the NRM (1) demonstrate suitable reliability and (2) assess the same behavioral objective. The level of agreement between these items will be determined when appropriate.

VALIDATION PROCEDURES FOR CRITERION REFERENCED MEASURES: MASTERY TESTS

IV. Percent correct: it is desirable than $\geq$ 80%[1] of students who have received the curriculum should pass each mastery test item.[2]

A. If $\geq$ 80% pass the item, go to V.

B. If < 80% pass the item, examine the item so as to discover possible deficiencies in the item (e.g., the item may be unrelated to the curriculum, the wording within the item may be misleading, or there may be errors in media).

(1) If these or other defects are apparent, momentarily discontinue further analyses of this item. Correct the apparent defects and readminister the items when it is possible to do so.

(2) If no defects are apparent, go to V.[3]

V. Determine the test-retest reliability for each item ((A+D)/N).

A. If test-retest reliability is suitably high, go to VI.

B. If test-retest reliability is unsuitably low, examine the item and the conditions under which the item was administered. Continue further analyses on this item only after the item is corrected and readministered.

---

1 As there are no firm guidelines for selecting 80% as the acceptable criterion, this figure can be adjusted. A downward or upward adjustment in the criterion level does not influence appreciably other aspects of the proposed decision model.

2 The percent correct figure for children who have not received the curriculum should be roughly equal to 1/number of options in the item (i.e., chance).

3 Even though < 80% pass an item, the item may have suitable reliability, thus ruling out one factor which may account for the low percent correct figure.

	P	F
P	A_{40}	B_{10}
F	C_{10}	D_{40}

VI. Validity: Mastery Test Items

There should be a high level of agreement between the level of achievement children demonstrate in the classroom (and observed by classroom personnel) and their performance on test items measuring achievement.

A. There should be a high level of agreement[4] ((A+D)/N) between the achievement level of students as estimated by their teacher and their performance on a mastery test item--provided that both refer to the same behavioral objective. The level of agreement will be determined for each mastery test item.

(1) If (A+D)/N is suitably high, go to VIB.

(2) If (A+D)/N is unsuitably low, try to determine possible reasons for the discrepancies in ratings and continue to VIB.

B. There should be a high level of agreement ((A+D)/N) between the achievement level of students as estimated by trained observers and their performance on a mastery test item--provided that both refer to the same behavioral objective. The level of agreement will be determined for each mastery test item.

(1) If (A+D)/N is suitably high, go to VIC.

(2) If (A+D)/N is unsuitably low, try to determine possible reasons for the discrepancies in ratings and continue to VIC.

C. There should be a high level of agreement between performance on an item from a mastery test and performance on an item from a NRM provided that the items from the Mastery Test and NRM (1) demonstrate suitable reliability and (2) assess the same behavioral objective. The level of agreement between these items will be determined when appropriate.

4

Teacher Ratings	Item Performance: P	Item Performance: F
P	A	B
F	C	D

REFERENCES

Bechtoldt, H. P. Construct validity: a critique. American Psychologist, 1959, 14, 619-629.

Cox, R. C. & Vargas, J. S. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, February 1966.

Cronbach, L. J. & Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.

Edmonston, L. P. A review of attempts to arrive at more suitable evaluation models: an introspective look. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April 1972.

Fisher, R. A. Statistical methods for research workers. New York: Hafner Publishing Co., Tenth Edition, 1948.

Gagne', R. M. & Paradise, N. E. Abilities and learning sets in knowledge acquisition. Psychological Monographs, 1961, 14, Whole No. 518.

Goodman, L. A. & Kruskal, W. H. Measures of association for cross classifications. American Statistical Association Journal, 1954, 49, 732-764.

Goodman, L. A. & Kruskal, W. H. Measures of association for cross classifications: III. Approximate sampling theory. Journal of the American Statistical Association, 1963, 58, 310-364.

Guttman, L. A generalized simplex for factor analysis. Psychometrika, 1955, 20, 173-192.

Kennedy, B. T. The role of criterion-referenced measures within the total evaluation process. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April 1972.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Oakland, T. An evaluation of available models for estimating the reliability and validity of criterion-referenced measures. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April 1972.

Roudabush, G. E. & Green, D. R. Some reliability problems in a criterion-referenced test. Paper presented at the Annual Meeting of the American Educational Research Association, New York, February 1971.

Scriven, M. The methodology of evaluation. Perspectives on curriculum evaluation. (Edited by Robert E. Stake) AERA Monograph Series on Curriculum Evaluation, No. 1, Chicago: Rand McNally, 1967, pp. 39-83.

Woodrow, H. The effects of practice on groups of differential initial ability. Journal of Experimental Psychology, 1938, 29, 268-278.